(12) LEVEL II

(9) Technical Paper 322

AD

(6)

AN EXPERIMENTAL INVESTIGATION OF THE MILITARY POLICE FIREARMS QUALIFICATION COURSE

(10) Frederick H. Steinheiser, Jr.

and

Kenneth I. Epstein

UNIT TRAINING & EVALUATION SYSTEMS TECHNICAL AREA

(14) ARI-TP-322

(16) 2Q762722A764

(12) 49p.

U. S. Army
Research Institute for the Behavioral and Social Sciences

(11) Sep[redacted]978

408 010

78 12 18 01

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

WILLIAM L. HAUSER
Colonel, US Army
Commander

JOSEPH ZEIDNER
Technical Director

NOTICES

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER Technical Paper 322	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) AN EXPERIMENTAL INVESTIGATION OF THE MILITARY POLICE FIREARMS QUALIFICATION COURSE		5. TYPE OF REPORT & PERIOD COVERED -- 6. PERFORMING ORG. REPORT NUMBER --
7. AUTHOR(s) Frederick H. Steinheiser, Jr., and Kenneth I. Epstein		8. CONTRACT OR GRANT NUMBER(s) --
9. PERFORMING ORGANIZATION NAME AND ADDRESS Army Research Institute for the Behavioral and Social Sciences (PERI-OU) 5001 Eisenhower Avenue, Alexandria, VA 22333		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q762722A764
11. CONTROLLING OFFICE NAME AND ADDRESS Deputy Chief of Staff for Personnel Washington, DC 20310		12. REPORT DATE September 1978 13. NUMBER OF PAGES 44
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) --		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE --

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

--

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Analysis of variance
Mastery learning
Classification errors
Criterion measurement

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Military Police Firearms Qualification Course (MPFQC) is a test of handgun marksmanship, composed of eight different "tables" or combinations of distance and position for firing at stationary targets. In this experiment, 237 male Military Police (MP) students fired three repetitions of the test over a 2-day period, for a total of 240 trials per student. The results of a variety of analyses converge on a single conclusion: Tables 1-4 and 5-8 comprise two dichotomous tests, the "difficult" and the "easy." Marginal students who fail the difficult tables may nonetheless pass the entire test OVER

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

78 12 18 0

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20.

by doing well on the easier tables. Classification errors showed that false positives and false negatives both averaged about 7%. False positives were perceived as being more serious than false negatives. The MPFQC works well as a training and testing instrument, although average scores must be interpreted with caution.

Technical Paper 322

AN EXPERIMENTAL INVESTIGATION OF THE MILITARY POLICE FIREARMS QUALIFICATION COURSE

Frederick H. Steinheiser, Jr.

and

Kenneth I. Epstein

UNIT TRAINING & EVALUATION SYSTEMS TECHNICAL AREA

Submitted as complete and
technically accurate, by:
Frank J. Harris
Technical Area Chief

Approved By:

A.H. Birnbaum, Acting Director
ORGANIZATIONS AND SYSTEMS
RESEARCH LABORATORY

Joseph Zeidner
TECHNICAL DIRECTOR

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

Office, Deputy Chief of Staff for Personnel
Department of the Army

September 1978

Army Project Number
2Q762722A764

Unit Training Standards
and Evaluation

ARI Research Reports and Technical Papers are intended for sponsors of R&D tasks and other research and military agencies. Any findings ready for implementation at the time of publication are presented in the latter part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

The research presented in this report was conducted under Project METTEST (Methodological Issues in Criterion-Referenced Testing), under the auspices of the Unit Training and Evaluation Systems Technical Area of the Army Research Institute for the Behavioral and Social Sciences (ARI), and under Army Project 2Q762722A764. The goal of Project METTEST is to develop quantitative methods for evaluating unit proficiency. The means for achieving this goal include basic research in test construction, measurement and decisionmaking models, and computer-programmable models for large-scale data analysis.

Related, ongoing programs within the UTES Technical Area include evaluation of small combat units under simulated battlefield conditions (REALTRAIN, ARTEP), qualification of tank gunnery crews and revision of table VIII (IDOC), and combat effectiveness evaluation by group decision making and board-game simulation (COTEAM, or Combat Operations Training Effectiveness Analysis).

Anticipated future research under Project METTEST includes the development of a computer-programed model for unit performance evaluation, application of extant quantitative models to performance data, and conducting "board-game" experiments to develop criterion-referenced methods for evaluating combat unit proficiency.

Research for this report was conducted with the cooperation of LTC P. Westin, MAJ D. King, and other representatives of the U.S. Army Military Police School (USAMPS), Department of Evaluation.

Joseph Zeidner

JOSEPH ZEIDNER
Technical Director

AN EXPERIMENTAL INVESTIGATION OF THE MILITARY POLICE FIREARMS QUALIFICATION COURSE

BRIEF

Requirement:

The educational decisionmaker often needs to determine if a student can perform a job at some specified level of acceptability. The level or "criterion" required for passing a job-related test affects the extent to which manpower quotas can be filled and the accuracy of classifying examinees as "masters" or "nonmasters." This experiment examined the ability of a hands-on performance test to accurately classify examinees, as well as to examine the structure of the test itself, in terms of score replicability.

Procedure:

The Military Police Firearms Qualification Course (MPFQC) consists of eight different "tables" (combinations of position and distance to target). Over a 2-day period, 237 MP students performed the test three times for a total of 240 shots per student. The data were analyzed statistically to determine how accurately the tables differentiate masters and nonmasters, and the results of changing the pass-fail criteria.

Findings:

The eight tables actually comprise two separate tests; tables 1-4 are relatively difficult and 5-8 are relatively easy. Marginal students who fail the difficult tables may nonetheless pass the entire test by doing well on the easier tables. Both false positive and false negative classification errors averaged about 7%. Military course experts perceived false positives to be about five times as serious as false negatives.

Utilization of Findings:

The MPFQC works well as a training and testing instrument, although average scores must be interpreted with caution. The technique of having experts equate some rate of false positives to a fixed amount of false negatives can be used to derive utilities for error rates. Easy items (tables) do little to separate masters from nonmasters, and the test may be shortened by eliminating such nondiscriminating items.

AN EXPERIMENTAL INVESTIGATION OF THE MILITARY POLICE FIREARMS QUALIFICATION COURSE

CONTENTS

LIST OF TABLES

LIST OF FIGURES

AN EXPERIMENTAL INVESTIGATION OF THE MILITARY POLICE FIREARMS QUALIFICATION COURSE

INTRODUCTION

The Military Police Firearms Qualification Course (MPFQC) is a practical course of instruction in police firearms training. Practical courses of fire include time limitations, protective barricades, firing from preferred and nonpreferred hand, the use of various bodily positions, and various distances to the target. The student receives live fire practice on a training range before "shooting for record" on the test or qualification course. The MPFQC is considered to be "the ultimate test of an MP's proficiency in police combat shooting (Army Training Circular 19-4, June 1975, Appendix F-1)." The purpose of the present experiment was to examine the properties of the MPFQC and to describe these properties to the Department of Evaluation, U.S. Army Military Police School (USAMPS).

The MPFQC may be considered an eight-item test, because there are eight different distance-position combinations, or "tables," to be fired. (See Figure 1 and the Method section.) The criterion for passing is at least a 70% hit rate over all eight tables combined.

In the most general sense, a test item is supposed to measure a student's ability or knowledge, as represented by that item. For performance tests such as the MPFQC test, there should be a strong causal relationship between answering test items correctly and being able to do the "real-world job" that those test items imply. Test items may also be used to measure the effectiveness of the training program if a group of students is split into two similar subgroups. For example, if group A students went through training program A and group B went through training program B, and if group A then scored twice as high on the same test as group B, it seems reasonable to conclude that the difference is not due to the students, but rather to the quality of the training programs. Finally, the effectiveness of a test or a testing procedure can be measured and evaluated. For example, if half the items on a test are so easy that 90% of the students get them right, and the other half are so difficult that 80% of the students get them wrong, then the test itself is failing to separate the "masters" from the "nonmasters" and could stand revision.

There are several requirements for describing and evaluating a testing procedure/program: (a) A large number of examinees should be tested under identical or highly similar conditions; (b) both the entire test and the same test items should be given to the same students several times; and (c) the students should have completed their training recently and successfully. In this experiment, designed to describe and

evaluate the MPFQC, each of 237 male MP students, after successfully completing practice firing with the .45 caliber automatic pistol, shot 10 rounds for each qualification table three times. The amount and quality of the data were therefore considered to be adequate for a variety of quantitative analyses which would provide a description and evaluation of the MPFQC.

Many applied testing programs such as the MPFQC face a variety of constraints.

1. The cost constraint requires a realistic upper bound to the number of test items or trials given to evaluate a student or crew. (This is especially obvious in tank crew live firing their table 8.) Time, manpower, and resource constraints are specific instances of the types of costs that are incurred in any testing program.

2. The quota constraint requires some percentage of students to pass to fill manpower vacancies.

3. The misclassification constraint specifies what percentage of students can be tolerated as either false positives or false negatives. This constraint can only be studied when students take the same test twice, as in this experiment.

4. Because not all items can be relevant to future tasks, any test can only sample from a vast storehouse of items; hence, the item sampling constraint requires that the most representative items be given. (It is not always easy to decide which are most representative and when to stop adding items. For example, possible MPFQC "items" of 30- and 10-meter distances have been omitted in the MPFQC presumably because they are less representative of real-life situations than those positions which it presently includes.)

5. Finally, a criterion constraint is often set as a matter of policy beyond the control of the trainers/testers. From a quality control perspective, stipulating "x"% or better on a test as a passing score should be justified to meet the previously mentioned constraints. For example, why set an "x"% criterion for overall test score instead of a "y"% on the easier items, and a "z"% ($z < y$) on the harder items? What are the misclassification consequences of an x% versus y% criterion? Who passes, who should have failed, and who fails who should have passed?

The remainder of this paper is divided into the following sections: The Method section describes how the experiment was conducted, the Results section presents the data in a variety of breakdowns, and each subsection explicitly states the hypothesis or question posed for the particular method of analysis used. The Discussion section integrates and explains

the various findings of the Results section, and the Summary and Conclusions section states the major findings and their implications.

METHOD

To obtain stable estimates of scores and parameters, and to reduce error variability, a large number of examinees and test trials were required. In all, 237 MP students were tested, using the .45 caliber handgun and firing at stationary silhouette targets. Each student fired a total of 240 shots (trials) over a period of 2 days. The 240 trials were divided into three repetitions of 80 trials each. Each group of 80 was divided into 10 shots for each of the eight "tables," or distance-position combinations, that are used as part of the standard operating procedure for the MPFQC. Each group of 10 shots was divided into 2 groups of 5 shots each, because the student had to reload after taking the first 5 shots to take the next.

The subjects for this experiment had completed practice firing of the .45 caliber automatic pistol on a practice range, 1 or 2 hours before participating in the present experiment on the qualification range. They had practiced each of the eight tables and had to have a practice score of at least 35 hits out of 50 shots before being allowed to "shoot for record" on the qualification range. (Ten shots were taken for tables 1 and 2, and five shots were taken for tables 3 through 8 on the practice range. If a student's first score was 34 or less, he repeated all eight tables until his score was 35 or more.)

Design

The experimental design is a completely crossed randomized block factorial (Kirk, 1968; Winer, 1971): A x B x C x D with A = 237 persons, B = 2 scores, C = 8 tables, and D = 3 repetitions.

Procedure

Groups of 25 students were tested on an outdoor firing range from November 1976 to early March 1977. Testing was not conducted on cold days when the temperature was below the mid-30's. The first trial occurred on Thursday mornings, the second on Thursday afternoons, and the third repetition on Friday mornings.

The order of firing for each of the repetitions was as follows.

Table	Range (meters)	Position	Maximum time (min., sec.)
1	35	lying prone	1-45
2	25	standing, no support, preferred hand	1-45
3	25	standing with support, weak hand	1-30
4	25	standing with support, preferred hand	1-30
5	15	standing, no support preferred hand	1-20
6	15	kneeling with support, left hand	1-20
7	15	kneeling with support, right hand	1-20
8	7	crouch	0-24

No feedback was available to a student until after he had fired all eight tables (80 trials). Visual sighting of bullet holes in the target was not possible by the student. Holes in the targets were covered with black tape by assistants after each score group of five shot trials while the students had their backs turned to the targets and were reloading for the next score group.

For each table, one five-round magazine was fired, and while the student reloaded, his score (from 0 to 5 hits) was recorded. After another five rounds, the score was recorded, and he then fired the next table. Thus there were two scores, from 0 to 5, for each table.

RESULTS

Figure 1 shows the distribution of the raw scores. The height of each bar indicates the number (or frequency) of students who obtained the score listed at the bottom. Recall that a perfect score here is 240 hits out of 240 shots. One person made a perfect score.

Figure 2 is a more condensed and comprehensible frequency-score distribution, in which scores have been grouped by fives. The peak of the distribution shows that the most commonly occurring scores range from 175 to 200. The dispersion to the right and left of the peak

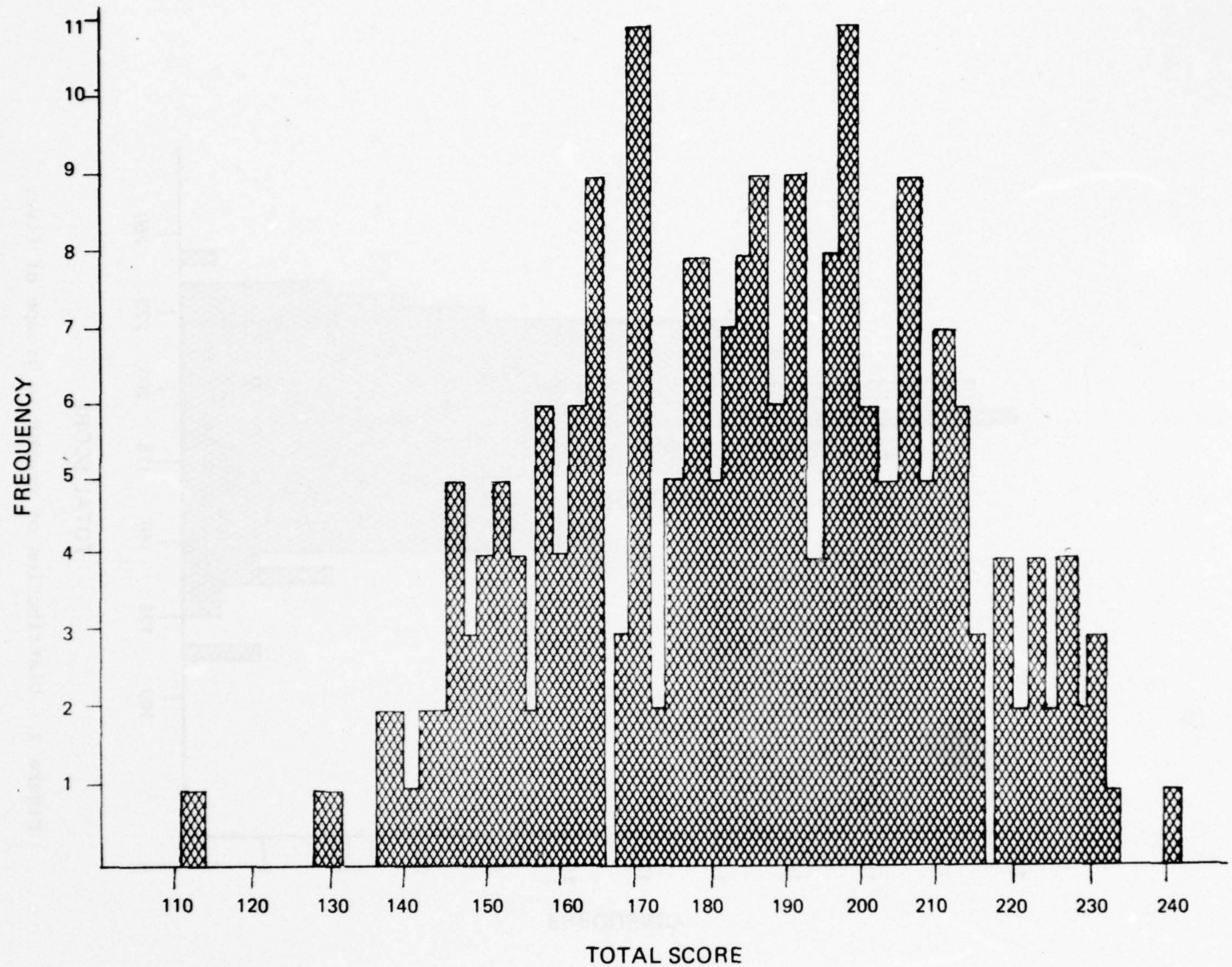

Figure 1. Raw score distribution.

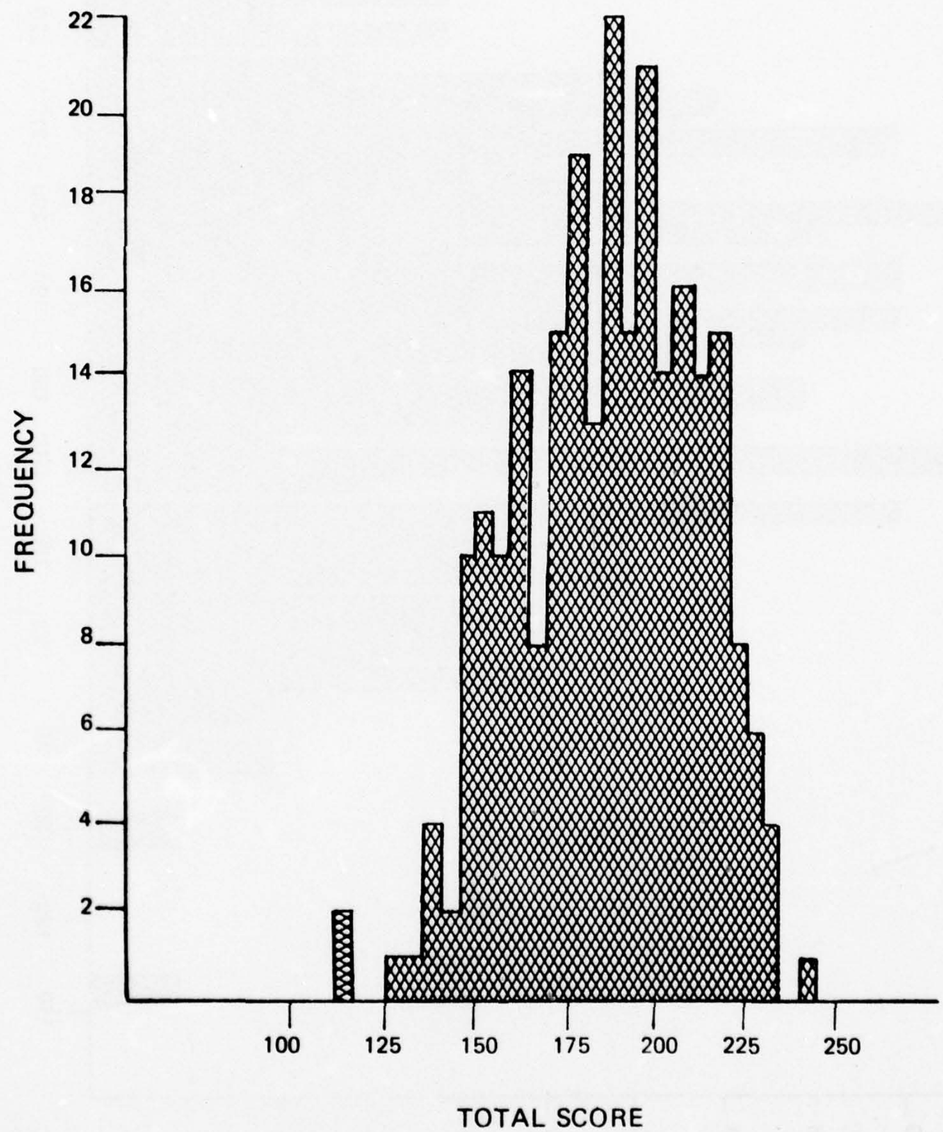

Figure 2. Distribution of scores in groups of five.

indicates that the higher or lower the score, the fewer were the students who obtained it.

Figure 3 shows the percentage of hits for each table and for each repetition. Note that tables 1 through 4 produced the lowest scores (from .572 to .679) and that tables 5 through 8 produced the highest scores (.837 to .975).

Figure 3 suggests that repetition produces some slight improvement, since the average Thursday-afternoon score (across all eight tables, based on 80 trials) was .5% higher than that for Thursday morning, and the Friday-morning scores were .8% higher than the Thursday-afternoon scores.

The next section examines the statistical significance of these scores, as a function of tables and repetitions.

Analysis of Variance

A four-factor completely crossed randomized block factorial analysis of variance was conducted, which has the following factors and levels: Students (A), $p = 237$; Scores (B), $q = 2$; Tables (C), $r = 8$; Repetitions (D), $s = 3$. The Student, Scores, and Repetitions factors were considered to be random, and Tables to be a fixed effect. The values for the expected mean squares in this design are presented in Appendix A, following the procedures in Kirk (1968, p. 7.10) and Winer (1971, p. 5.14). Quasi-F ratios were derived following the procedures recommended by these authors. (See Appendix B.)

The results of this analysis of variance are presented in Table 1. The following statistically significant main effects emerged: Persons, $F(272, 705) = 3.93$, $p < .001$; Scores, $F(1, 25) = 5.96$, $p \leq .025$; Tables, $F(7, 42) = 79.11$, $p < .001$; Repetitions, $F(2, 87) = 12.55$, $p < .001$. The following significant interactions also emerged: Persons x Tables, $F(3100, 4695) = 1.33$, $p < .001$; Persons x Repetitions, $F(472, 472) = 2.52$, $p < .001$; Scores x Tables, $F(13, 75) = 1.94$, $p \leq .05$; Tables x Repetitions, $F(4, 99) = 2.82$, $p \leq .03$; Persons x Scores x Tables, $F(1652, 3304) = 1.11$, $p \leq .05$; Persons x Tables x Repetitions, $F(3304, 3304) = 1.38$, $p < .001$. Thus, only two double interactions and one triple interaction failed to achieve statistical significance.

Interpretation of these results will be aided by inspection of Figure 3. Many "post-hoc" comparisons could be made among various parsings of the data matrix. For example, were fewer hits made on table 1 on Thursday morning than on Friday morning, or on table 2 on Thursday afternoon than on Friday morning? The most pertinent comparison, however, deals with the tables themselves, regardless of the particular test repetition in which they were fired. Some tables were clearly easier than others, with more hits scored on tables 5 through 8 than on tables

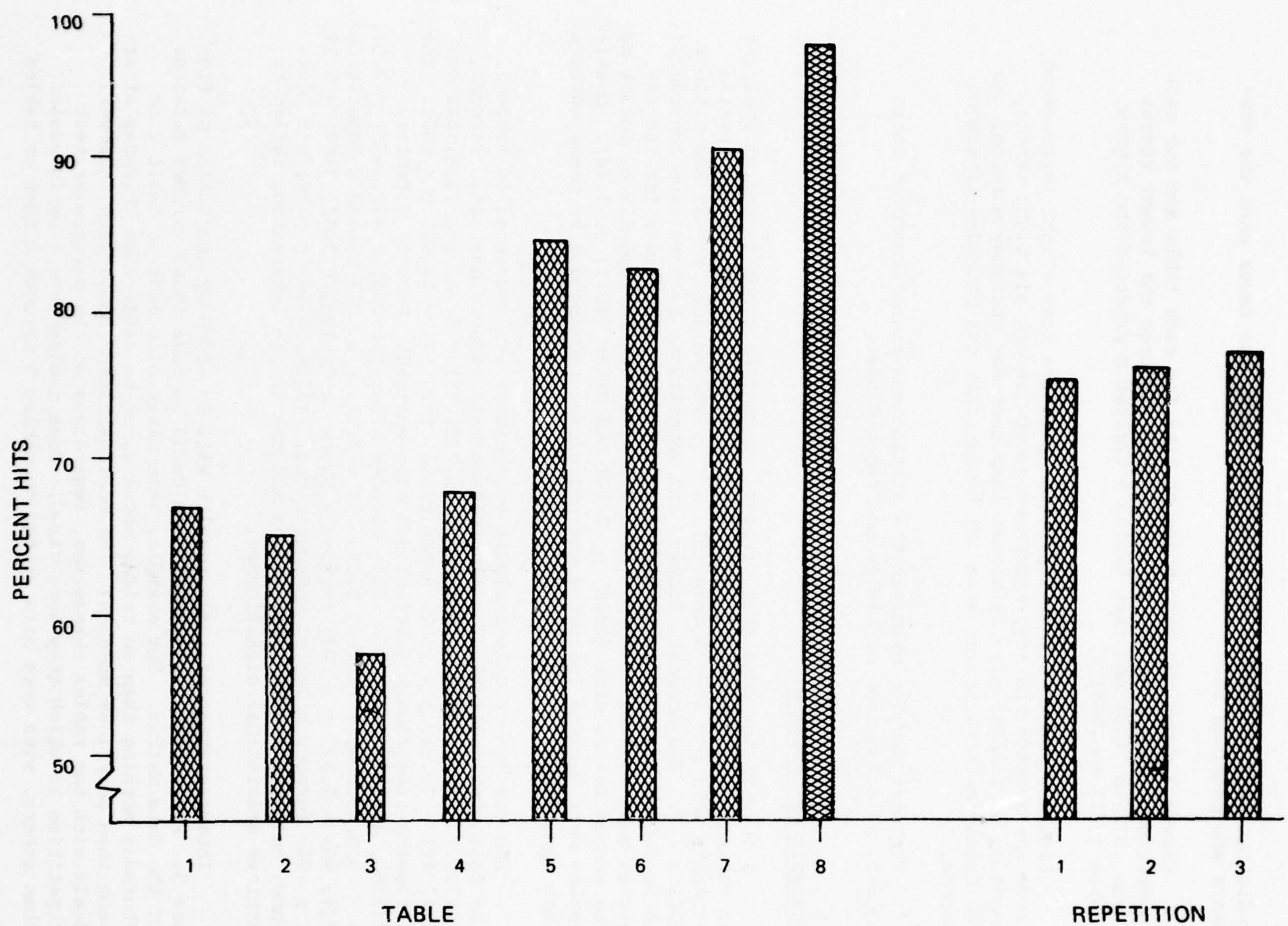

Figure 3. Average scores (expressed as percentage of hits) as a function of tables and repetitions.

Table 1

Analysis of Variance for Completely Crossed, Mixed Model:
A, B, D Random, C Fixed

Source	df	M.S.	F[a]	df[b] Numerator	Denominator
A (Persons)	236	12.80	3.93****	272	705
B (Scores)	1	7.70	5.96**	1	25
C (Tables)	7	732.71	79.11****	7	142
D (Repetitions)	2	34.75	12.55****	2	87
AB	236	1.05	1.09	236	472
AC	1652	1.90	1.33****	3100	4695
AD	472	2.45	2.52****	472	472
BC	7	2.26	1.94*	13	75
BD	2	.40	.41	2	472
CD	14	4.31	2.82***	4	99
ABC	1652	.91	1.11*	1652	3304
ABD	472	.97	untestable		
ACD	3304	1.14	1.38****	3304	3304
BCD	14	.68	.83	14	3304
ABCD	3304	.82	untestable		

[a]See Appendix B for derivation of "quasi-F Ratios."

[b]See Appendix C for derivation of degrees of freedom (df).

*p < .05.
**p < .025.
***p < .01.
****p < .001.

1 through 4. The difference between the average of these two groups was shown to be statistically significant, using a Tukey test (Winer, 1971, p. 3.9; Kirk, 1968, p. 8.7), with the value being $Q(8, 42) = 36.93$, $p < .001$. Furthermore, the easiest of the four most difficult tables (4) was also shown statistically to be significantly more difficult than the most difficult of the four easiest tables (6), using the same type of Tukey test: $Q(8, 42) = 23.24$, $p < .001$. These results strongly imply that there are two "types" of tables, the easy and the difficult, and that this is a genuine, not statistically artificial, dichotomy. Computational details for the Tukey post-hoc comparison procedure may be found in Appendix D.

Magnitudes of Effects

Because of the large sample size (237 subjects), there is considerable power to reject null hypotheses and to obtain significant treatment effects in the analysis of variance. Effects shown to be statistically significant may actually account for a rather minuscule portion of the total score variance. Therefore, the relative magnitudes of the experimental effects (also called proportions of variance) were calculated (Winer, 1971; Cronbach et al., 1972; Dodd & Schultz, 1973). (See Appendix E.) As in the previous analysis of variance, the Tables factor was considered a fixed effect, and Persons, Scores, and Repetitions were considered to be random factors. These results are presented in Table 2, where it may be seen that the largest effect (other than that due to random error) was due to the Tables variable, with a share of 23% of the total magnitude. The effect due to Persons, reflecting individual differences among the students, reached nearly 10%. Several interaction terms, in which Tables was a factor, accounted for about 6% to 7%.

Note that the analysis of variance effect due to Repetitions in Table 1 was statistically significant, whereas according to Table 2, Repetitions contributed an effect worth only about .4%. This apparent discrepancy between the two methods of analysis is due to the large number of subjects, which produces a large value for degrees of freedom and allows small F ratios to achieve statistical significance. Thus, the values in Table 2 act as a sort of check upon the significance levels in Table 1. The effect due to Repetitions therefore reveals a slight, but probably inconsequential, learning effect. A similar line of reasoning holds true for interpreting the effect due to the Scores variable in each of these tables.

Classification Accuracy

Each student fired the MPFQC three times to allow comparisons to be made among the three repetitions. A specific question which this design allows to be answered is this: If a student passed table "x" on the first repetition, did he also pass it on the second? If he passed

it on the second, did he also pass it on the third? If he passed it on the first, did he also pass it on the third?

Table 2

Magnitudes of Effects or (Proportion of Total Score Variance Accounted for by a Given Source of Variance)

Effect (source of variance)	Proportion of variance, or magnitude of effect
A (Persons)	.1027
B (Scores)	.0006
C (Tables)	.2454
D (Repetitions)	.0041
AB	.0017
AC	.0536
AD	.0444
BC	.0007
BD	.0
CD	.0032
ABC	.0144
ABD	.0582
ACD	.0769
BCD	.0
ABCD (error)	.3939

Ideally, if a student fails a table once, then he should not be able to pass it any other time. But suppose that he passes it once and fails it once: Then what is the final decision--does he pass or fail that table? Or, suppose that his total score on the first repetition is below 70%, and on the second is above 70%. Had the pass-fail decision been based only upon the first score, he would fail. Actually, many examinees passed the test one time but failed it another. The significant aspect of conducting the same test three times is that false positive and false negative error rates can be specified for the MPFQC. The values in Table 3 are these error rates.

Some examples follow, to interpret the values in this table. The general scheme is that a cell entry in the left column shows how many students (and what percentage of the total sample of 237 that number represents) passed a particular table but failed an entire test (either Repetition 1, 2, or 3). A cell entry in the right (false negative) column

shows how many and what percentage failed a particular table but passed the entire 80-trial test.

Table 3

Classification Errors for 70% Criterion

Table	False positive Number	False positive % of total N	False negative Number	False negative % of total N
		Repetition (test) 1		
1	17	7.2	63	26.6
2	9	3.8	78	32.9
3	8	3.4	88	37.1
4	12	5.1	72	30.4
5	36	15.2	5	2.1
6	33	13.9	18	7.6
7	48	20.3	2	.8
8	60	25.3	1	.4
		Repetition (test) 2		
1	10	4.2	50	21.1
2	11	4.6	60	25.3
3	5	2.1	102	43.0
4	12	5.1	52	21.9
5	37	15.6	8	3.4
6	29	12.2	8	3.4
7	52	21.9	8	3.4
8	59	24.9	1	.4
		Repetition (test) 3		
1	18	7.6	59	24.9
2	22	9.3	58	24.5
3	15	6.3	81	34.2
4	25	10.5	45	19.0
5	51	21.5	9	3.8
6	46	19.4	12	5.1
7	54	22.8	1	.4
8	60	25.3	0	0.0

Table 3 (Continued)

Table	False positive Number	False positive % of total N	False negative Number	False negative % of total N
		Total score		
1	12	5.1	15	6.3
2	7	3.0	13	5.5
3	29	12.2	20	8.4
		Between repetitions (tests)		
Comparison				
Test 1 vs. Test 2	25	10.5	22	9.3
Test 1 vs. Test 3	27	11.4	39	16.5
Test 2 vs. Test 3	24	10.1	39	16.5

Note. Classification is from table score to test score, test score to total score, and test score to test score.

N = 237.

Consider first false positives: For Test 1, 17 out of 237 students (7.2%) got seven or more hits out of 10 shots for table 1, but got fewer than 56 hits on the entire 80-trial test. (70% of 10 = 7; 70% of 80 = 56.) Note that for all three repetitions, the false positive rate for tables 1 through 4 tends to be quite low, less than 10% (except for table 4 in Test 3). This finding makes sense, because tables 1 through 4 are quite difficult (recall the percentage of hits in Figure 3), and it would be quite unusual to find that many people can pass the difficult items and yet fail on the overall test. But for the easier tables (5 through 8), note that the false positive rate ranges from a low of 12.2% to a high of 25.3%. This means that from 12.2% to 25.3% of the students who passed on tables 5 through 8 actually failed the entire test, regardless of which repetition it was.

These findings are reversed for false negative error rates. For the more difficult tables (1 through 4), anywhere from 19% to 43% of the students failed on one of these tables (fewer than seven hits out of each 10 trials), yet passed the entire test (56 or more hits out of 80 trials). This finding is reasonable because tables 1 through 4 are more difficult. The false negative error rate for tables 5 through 8, on the other hand, ranges from 7.6% to zero, which means that relatively few people failed these tables yet passed the entire test.

When each 80-trial test (repetition) is compared against the total test score (based upon 240 trials), a false positive error would occur when a student gets 56 or more hits on one test, but fewer than 70% of 240 (168) for all three tests. As shown in Table 3, this error rate ranges from 3% to 12.2%.

Similarly, if a student failed a particular test (fewer than 56 hits) but got at least 168 hits on all three tests combined, then a false negative error would be made. This error rate ranges from 5.5% to 8.4%.

Thus, the average false positive error rate is a bit less than 7% and the average false negative error rate is a bit more than 7%. The interpretation and implication of these values will be developed in the Discussion section.

The last observation to be made about miscl... ifications in Table 3 is that 10.5% (25 students) passed Test 1 but failed Test 2; 11.4% passed Test 1 but failed Test 3; and 10.1% passed Test 2 but failed Test 3. Thus, even if a student passes the test once, there is about a 10% chance that he will fail it the next time.

The false negative column reveals that 9.3% failed Test 1 but passed Test 2; 16.5% failed Test 1 but passed Test 3; and 16.5% failed Test 2 but passed Test 3. Hence, a student who failed a test the first time has about a 14% chance of passing it the next time.

Relative Value of Responses to Test Items

We can compute the "information content" in a response to each of the tables to find out the relative value of the tables. The formula is simply the probability of getting a hit on a table times the probability of not getting a hit on a table; symbolically, $p(1 - p)$. We will now simply use the probability values from the three right-hand columns in Appendix F. The values shown are for an average student (mean); a good student whose score was one standard deviation above the average; and a poor student whose score was one standard deviation below the mean. Small values shown for a given table indicate it yielded "information content." For example, considering table 8 as one item and table 1 as another, we find that the value of responses to table 1 is about 20 times as great as for table 8. The general pattern is that tables 1 through 4 are fairly informative, and tables 5 through 8 are relatively uninformative, about how well a student performs; they thus comprise two very different types of test items, those that are relatively informative (tables 1 through 4) and those that are not (tables 5 through 8).

Relative Utility of Misclassifications

Because 100% correct classification accuracy is impossible to achieve, it is helpful to determine how "bad" it is to make false positive and false negative errors. Specifically, if a classification error will be made, is it "better" to fail a truly qualified student or to pass a truly unqualified student? (Inspection of Table 3 shows that both types of errors occurred in all phases of the MPFQC.)

The consequence of passing an unqualified student is that the student will have use of a weapon which he might misuse in an emergency. The consequences of failing a qualified student are failure to meet a quota of manpower needs and the need for yet more time and expense to retrain and retest that student. Thus, the relative costs of misclassification are largely subjective and transcend the specific (MPFQC) training and testing program.

An attempt was made to determine the relative tradeoff between the two types of errors. If it is just as "bad" to commit a false positive error as a false negative error, then an MPFQC content expert should not have any strong preference in choosing, because the alternatives are equally bad. Intuitively, it would seem worse to pass an unqualified student than to fail a qualified student, because these errors have different consequences.

Six experts in the MPFQC were asked the following question: "Given that "x" number of unqualified students are going to be incorrectly passed (as false positives), then how bad a mistake do you think that is when compared to failing some number ("y") of qualified students?"

Each expert was provided with a scoresheet, with varying numbers of unqualified students passing listed down the left column, ranging from 1 to 4. For each number of false positives, the expert was asked to select a number for failing qualified students in the right column, so that the subjective values in the columns would be equal from their point of view. For example, if one false positive is assumed, and an expert marks a six in the false negative column, then for that expert, passing one unqualified student is about as bad as failing six qualified students.

Figure 4 shows the results of this study, equating the value of the average number of false negatives selected by six experts to the designated number of false positives. There is a highly consistent relationship; one false positive is about as bad as six false negatives for the four values of false positives that were given.

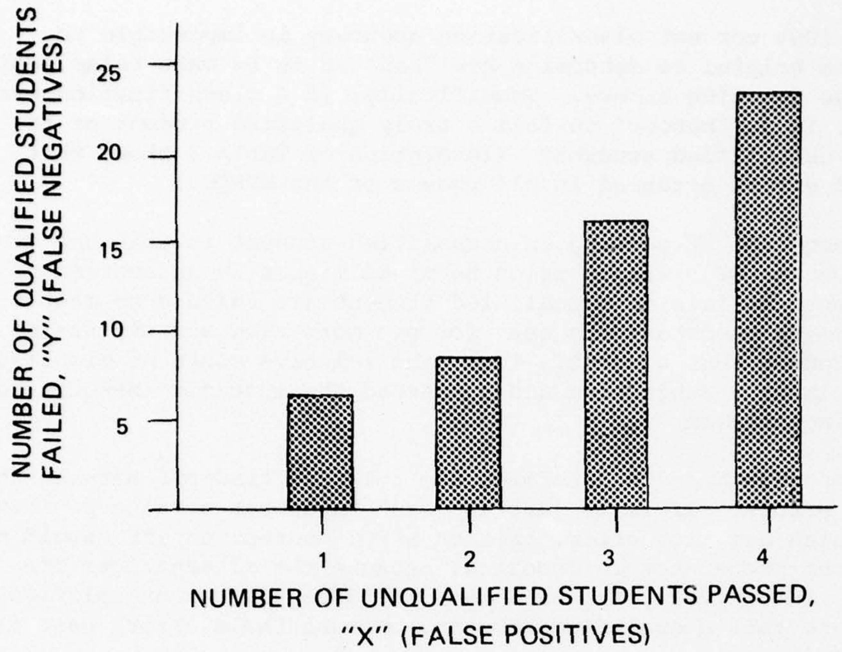

Figure 4. Relative costs for committing a false positive versus a false negative classification error.

DISCUSSION

One important aspect of a test designed to classify examinees as masters and nonmasters is the accuracy with which test classifications are made. Because an examinee's true ability can never be precisely measured, we need to obtain realistically close approximations. Three approximations were available from the design and data collected in this experiment.

First, classifications made on the basis of a table score alone were compared with those made on all 80 trials of one repetition, and these comparisons were repeated for each of the three repetitions. (See Figures 5 and 6.) When comparing classifications based upon the 10-shot table scores to those from the 80-shot repetition, it may be seen that for tables 1 through 4 the false positive error rate is quite low (Figure 5), but that the false negative error rate is quite high (Figure 6). This means that very few students who failed the 80-shot test were able to pass the 10-shot tables 1 through 4. Also, a rather large number of students who passed the 80-shot test also failed the 10-shot tables 1 through 4 (Figure 6).

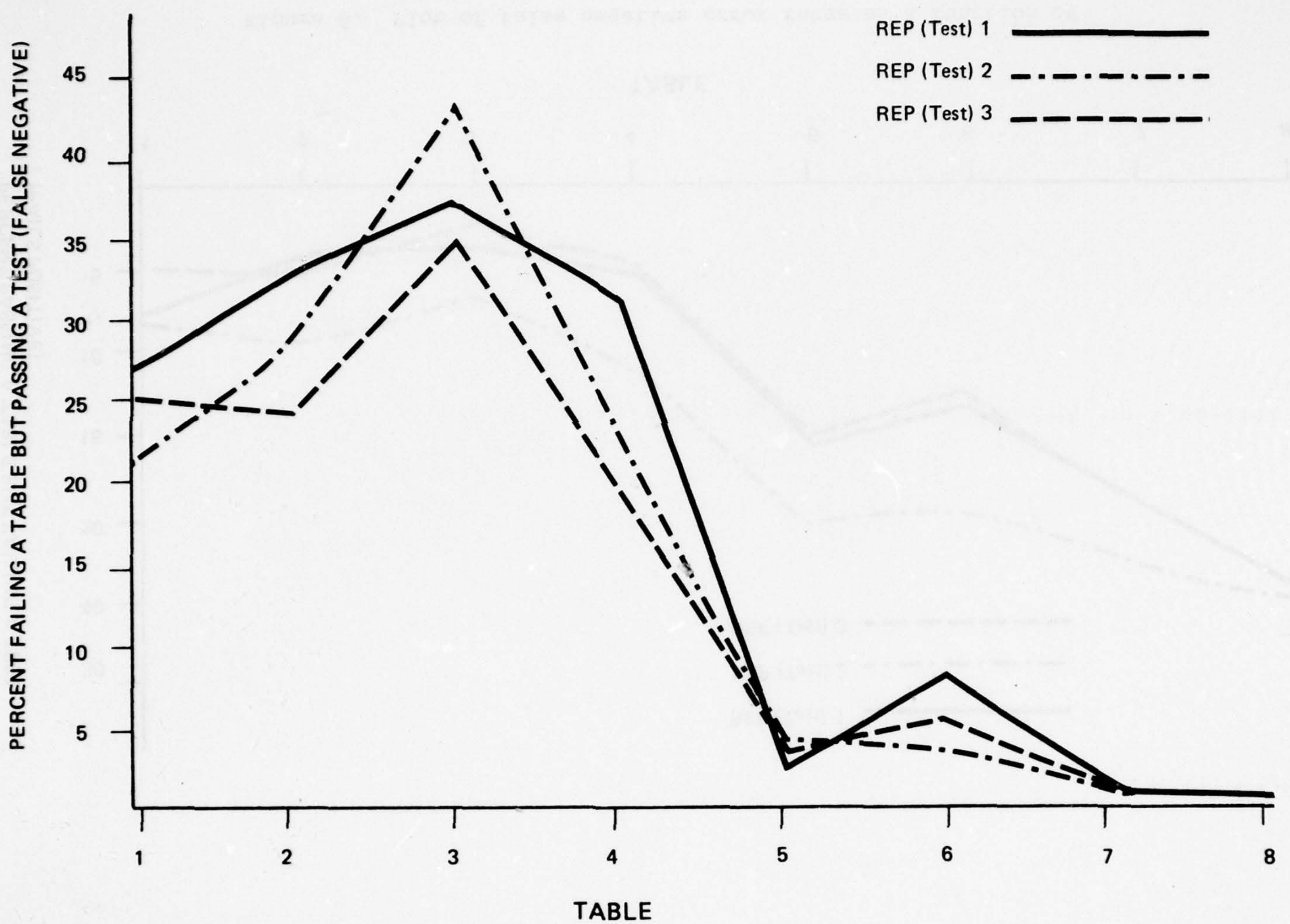

Figure 5. Plot of false positive error rates as a function of tables and repetitions of the test.

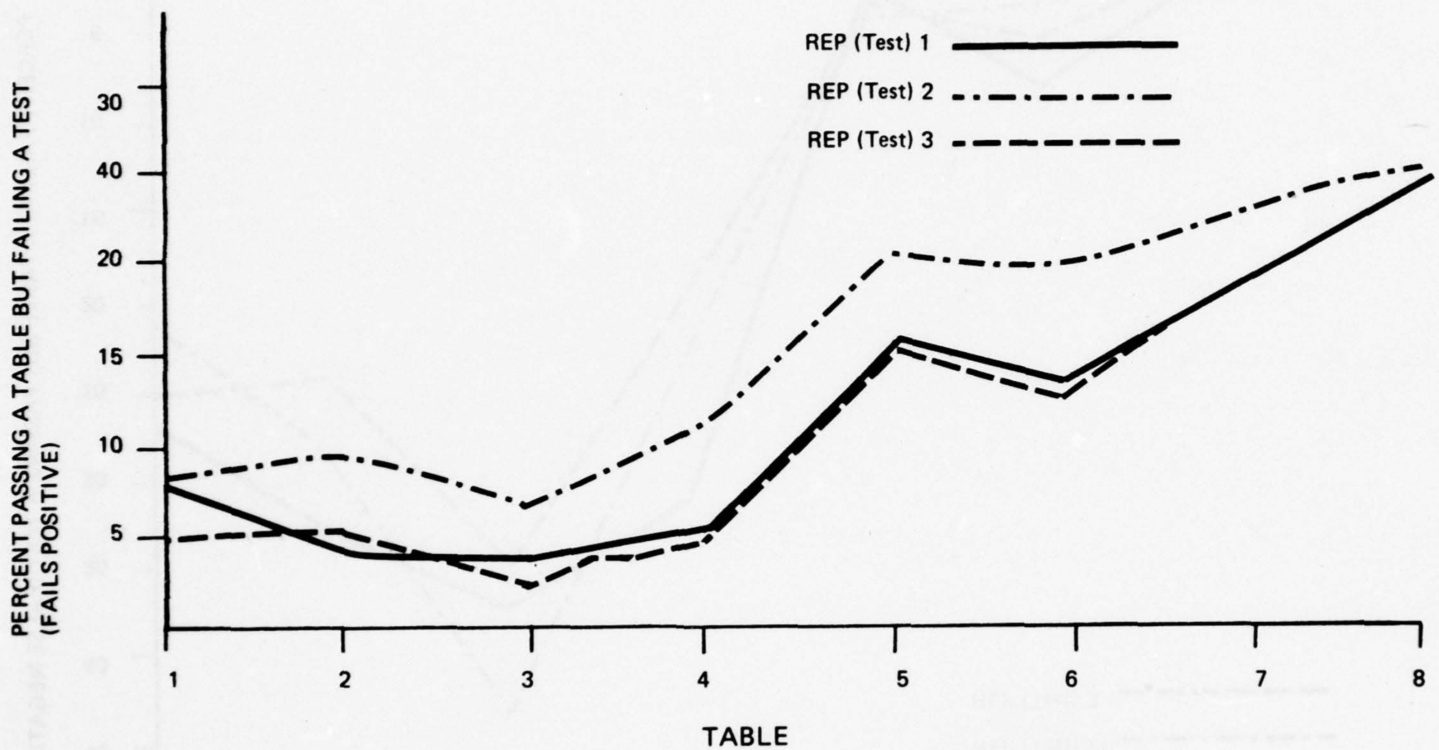

Figure 6. Plot of false negative error rates as a function of tables and repetitions of the test.

Since the first four tables were the most difficult (Figure 3), it is unlikely that a student would do well on them who did not do well overall. This is a desirable feature of a test. On the other hand, some students who did well enough overall were not able to meet the criterion of the more difficult tables.

For tables 5 through 8 these results are reversed. False positive error rates are relatively high (Figure 5) and false negative error rates are quite low (Figure 3). Since tables 5 through 8 are much easier than tables 1 through 4 (Figure 3), it is not very surprising that some students who did not do well overall were successful on the easier tables, and thus became false positives. Only a very few of those who did well on the overall test (70% or more hits) had trouble with tables 5 through 8 (less than 70%), leading to a low false negative error rate.

Second, in comparing the results of each of the 80-shot tests, we find relatively low error rates in all cases. This is a desirable outcome, suggesting that the 80-item test seems to be working. Because a test represents only a sample of all possible observations, classification errors occasionally occur due to random uncontrollable effects. Since no particular pattern of errors is apparent across the three repetitions, the three 80-item tests are operating in identical manner.

Third, a comparison of the classifications based upon each of the 80-trial test scores with the classification based on all 240 trials (three repetitions) was made. If classifications were perfect, the comparisons should show that *all were consistent*. As Table 3 indicates, the only obvious inconsistency is the low false positive rate for Tests 1 and 2 and the large false positive rate for Test 3. A plausible interpretation of this discrepancy is that students had learned enough through practice on Tests 1 and 2 to do well on Friday morning, yet their overall score from all three tests was not sufficiently boosted by passing Test 3 to classify them as masters of the total test.

The results of the analysis of variance show that the Persons, Scores, Tables, and Repetitions variables all had statistically significant effects. But as the subsequent "post-hoc" and "magnitude of effects" analyses demonstrated, the overwhelming effect of most practical significance was due to the Tables variable. The MPFQC is actually composed of two types of "test items," the relatively difficult (tables 1 through 4) and the relatively easy (tables 5 through 8).

This "two-test" interpretation is further substantiated by inspection of Table 4. Here, it may be noted that relatively little information is gained about a student's ability by testing him with tables 5 through 8, whereas the information gained by using tables 1 through 4 is about twice that from tables 5 through 8. This finding makes intuitive sense also, since extremely easy items would allow many nonmasters as well as masters to score many hits. Also, extremely difficult items would prevent many masters as well as nonmasters from scoring many hits.

In the present MPFQC, tables 7 and 8 are particularly easy, and do a minimal job of discriminating between masters and nonmasters. In this experiment, they essentially added 18 to 20 free and easy points to everyone's score for each repetition.

Table 4

Information Value of the Eight Tables as a Function of Average, Above Average, and Below Average Marksmanship Ability

Table	Average score: Logistic estimate[a]	Average score: Raw score[b]	One S.D. above average	One S.D. below average
1	.215	.223	.146	.250
2	.218	.230	.153	.250
3	.243	.246	.189	.240
4	.210	.225	.141	.250
5	.109	.128	.056	.177
6	.120	.137	.065	.200
7	.065	.082	.044	.120
8	.020	.030	.010	.040

[a] See Appendix G for the derivation of these logistically estimated probabilities.

[b] Raw score values may be obtained by referral to Figure 3. Thus, for table 1, $p = .670$, $1 - p = .330$, and $p(1 - p) = .223$.

The information about the relative costs of making false positive and false negative classification errors provided by the six MPFQC experts is important for interpreting the practical significance of the error rates listed in Table 3. The experimental finding was an average of about 7% false positives and about 7% false negatives. However, the experts claimed that a false positive was about as "bad" as six false negatives; that is, it would be worthwhile to fail (continue to retrain and retest) six marginal students just to prevent one from being incorrectly passed through the MPFQC as a false positive. If the experts' value judgments are accurate interpretations of the costs of these two types of misclassification, then the false positive error rate should be reduced to about 1%. This may be realistically unachievable, but the discrepancy serves to point out that a test may not be able to achieve precisely what its designers intend it to. The dilemma is by no means unique to the USAMPS!

Several additional analyses were made to illustrate the impact of modifications upon the criterion for passing and the actual structure of the MPFQC. The first analysis addresses how misclassification rates change as a function of the criterion for passing. That is, suppose that a higher (or lower) criterion were set. The entries in Table 5 reflect such changes in misclassification rates as systematically decreasing the criterion to 60% and increasing it to 80%. For example, with a 70% criterion, Table 3 reveals that 5.1% of the students who passed Test 1 did not pass the entire (total) 240 trial test. In comparison, Table 5 shows that with a 65% criterion, only 3% of the students who passed Test 1 did not get a passing total score. In general, the false positive rate decreases with a lower criterion, while the false negative rate shows no clear trend. What this means is (a) that the number of students who pass on a given test repetition, yet fail on the basis of their overall (Tests 1 + 2 + 3) score, tends to decrease as the criterion becomes more lax; and (b) that the number of students who fail a given test repetition, yet pass on the basis of their total score, is independent of the criterion level, for the range of 60% to 80%.

As can be seen in Figure 3, table 8 is so easy that 97.5% of the students mastered it. While it is undoubtedly worthwhile to train students to fire from a distance of seven meters, perhaps it is an unnecessary use of time, manpower, and material to test them on so easy (nondiscriminating) an item. What would the pattern of misclassification be, therefore, if table 8 were eliminated from this data set?

The results of deleting table 8 are shown in Table 6. It may be easily seen that the false positive rate has consistently decreased (compare Table 6 to Table 5). Thus, many false positives are due to passing scores on the easy table 8, with below-criterion scores on the total test. If the easy "item" is deleted, then fewer people pass who ought not to.

The effect of deleting easy table 8 on the false negative rate is to nearly double it, as can be seen by comparing the right-hand columns of Tables 5 and 6. This means that not many more students would fail who would otherwise (with the help of easy table 8) have passed.

If table 8 were eliminated from the MPFQC, and if the minimal passing score were kept at 70%, then fewer students would pass. Using the data for a 70% criterion, Table 5 shows that a total of 48 false positives occurred over the three test repetitions; Table 6 shows that a total of 30 false positives occurred when table 8 was omitted. The average percents are 6.7% and 4.3%, respectively. Given these data, the net failure rate would increase by about 2.4%, if table 8 were deleted.

Because it can be seen from Figure 3 that tables 1 through 4 are more difficult than tables 5 through 8, we might consider eliminating the hardest table and the easiest table. Table 7 shows how the misclassification rates change after tables 4 and 8 have been eliminated. There is a slight tendency for false negatives to increase.

Table 5

Classification Errors from Test Score to Total Score as a Function of Criterion

Criterion (%)	Repetition (test)	False positive Number	False positive % of total N	False negative Number	False negative % of total N
60	1	3	1.3	17	7.2
	2	3	1.3	16	6.8
	3	6	2.5	8	3.4
62.5	1	6	2.5	18	7.6
	2	6	2.5	18	7.6
	3	10	4.2	7	3.0
65	1	7	3.0	16	6.8
	2	4	1.7	20	8.4
	3	20	8.4	10	4.2
67.5	1	10	4.2	17	7.0
	2	7	3.0	19	8.0
	3	24	10.1	17	7.2
70	1	12	5.1	15	6.3
	2	7	3.0	13	5.5
	3	29	12.2	20	8.4
72.5	1	24	10.1	23	9.7
	2	15	6.3	23	9.7
	3	27	11.4	12	5.1
75	1	20	8.4	30	12.7
	2	21	8.9	13	5.5
	3	29	12.2	8	3.4
77.5	1	13	5.5	26	11.0
	2	21	8.9	14	5.9
	3	34	14.3	10	4.2
80	1	16	16.8	15	6.3
	2	23	9.7	11	4.6
	3	40	16.9	11	4.6

Table 6

Classification Errors as a Function of Criterion

Criterion[a] (%)	Test minus table 8	False positives Number	False positives % of total N	False negatives Number	False negatives % of total N
60 (60)	1	3	1.3	32	13.5
	2	1	.4	34	14.3
	3	3	1.3	13	5.5
62.5 (61.4)	1	3	1.3	33	13.9
	2	4	1.7	37	15.6
	3	8	3.4	25	10.5
65 (65.7)	1	2	.8	31	13.1
	2	3	1.3	35	14.8
	3	17	7.2	33	13.9
67.5 (67.1)	1	4	1.7	28	11.8
	2	3	1.3	29	12.2
	3	19	8.0	30	12.7
70 (70)	1	4	1.7	35	14.8
	2	5	2.1	23	9.7
	3	21	8.9	28	11.8
72.5 (72.9)	1	2	.8	16	6.8
	2	8	3.4	48	20.3
	3	8	3.4	29	12.2
75 (74.3)	1	10	4.2	41	17.3
	2	14	5.9	23	9.7
	3	23	9.7	15	6.3
77.5 (77.1)	1	10	4.2	31	13.1
	2	12	5.1	22	9.3
	3	28	11.8	20	8.4
80 (80)	1	11	4.6	24	10.1
	2	14	5.9	23	9.7
	3	26	11.0	21	8.9

Note. Classification is from Test minus table 8 to Total Score.

[a]Criterion % refers to total score out of 240 trials. The percentages for Test minus table 8 data in this table are approximate, because only 70, not 80, trials were taken. Actual percentage is given in parentheses.

Table 7

Classification Errors as a Function of Criterion

Criterion[a] (%)	Test minus tables 4 and 8	False positives Number	False positives % of total N	False negatives Number	False negatives % of total N
60 (60)	1	3	1.3	28	11.8
	2	2	.8	25	10.5
	3	4	1.7	14	5.9
62.5 (61.7)	1	4	1.7	22	9.3
	2	6	2.5	27	11.4
	3	10	4.2	13	5.5
65 (65)	1	5	2.1	22	9.3
	2	3	1.3	25	10.5
	3	20	8.4	21	8.9
67.5 (66.7)	1	7	3.0	20	8.4
	2	5	2.1	23	9.7
	3	23	9.7	20	8.4
70 (70)	1	6	2.5	25	10.5
	2	7	3.0	16	16.8
	3	22	9.3	25	10.5
72.5 (71.7)	1	16	6.8	22	9.3
	2	15	6.3	32	13.5
	3	15	6.3	19	8.0
75 (75)	1	13	5.5	35	14.8
	2	12	5.1	20	8.4
	3	22	9.3	22	9.3
77.5 (76.7)	1	14	5.9	23	9.7
	2	19	8.0	19	8.0
	3	28	11.8	16	6.8
80 (80)	1	12	5.1	24	10.1
	2	16	6.8	21	8.9
	3	28	11.8	17	7.2

Note. Classification is from Test minus tables 4 and 8 to Total Score.

[a]Criterion % refers to total score out of 240 trials. The percentages for Test minus tables 4 and 8 in this table are approximate, since only 60 instead of 80 trials are represented. Actual percentage is given in parentheses.

SUMMARY AND CONCLUSIONS

The experimental data collected on the MPFQC represent an extremely rich base for studying the operational characteristics of a carefully designed and executed performance test. This report described one approach to analyzing the data. A variety of statistical analyses were conducted on student markmanship performance scores, all of which tended to point toward consistent conclusions. More exhaustive, in-depth analyses are anticipated for future basic research; however, based on the present analyses, a rather clear view of how the test operates did emerge.

Taken as a whole, the test seems to be a relatively reliable index of marksmanship skills. That is, student performance tends to remain stable over time when the eight-table test is considered as an entity. However, within the test, a "difficult" subtest comprising tables 1 through 4, and an "easy" subtest comprising tables 5 through 8 are very distinct. This dichotomy in the test serves to identify two distinct classes of marksmen: those who score well on all tables from 35 to 7 meters, and those who are able to score well only at close distances from 15 to 7 meters (tables 5 through 8).

There are two practical implications of these findings. First, because nearly all students score well at close range, little information is gained (from a classification point of view) in testing on these tables. The longer range tables separate the truly proficient marksmen from those who are marginal. Second, marginal scores on the difficult tables combined with good scores on the easy tables can result in a marginal marksman being classified as "proficient." This probably accounts for the rather high false positive error rate.

The purpose of this experiment (and the interpretation of the results) is not to recommend that the MPFQC be drastically changed. It is the authors' understanding that the test serves training as well as testing purposes, and that the tables were designed to reflect realistic problems that MP's typically face in their assignments. From that point of view, the test is very satisfactory. We also conclude that the test works well as an average measure of marksmanship skill. However, it is important to note that this average measure has distinct parts; that is, the average score is not necessarily a good indicator of what a student can do on a given table. In fact, his performance on tables 1 through 4 may well be less than his average score, and his performance on tables 5 through 8 may well be much higher.

In sum, the results of this experiment provide one type of information which, if combined with others, such as training value of the test, manpower requirements, and resource constraints, could point the way to modifications in the MPFQC.

REFERENCES

Cronbach, L., Glesner, G. C., Nanda, H., & Rajaratnam, N. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley, 1972.

Dodd, D. H., & Schultz, R. F. Computational Procedures for Estimating Magnitude of Effect for Some Analysis of Variance Designs. Psychological Bulletin, 1973, 79, 392-395.

Kirk, R. E. Experimental Design: Procedures for the Behavioral Sciences. Belmont, Calif.: Brooks/Cole Publishing, 1968.

Winer, B. J. Statistical Principles in Experimental Design (2nd ed.). New York: McGraw-Hill, 1971.

Wright, B. D. Solving Measurement Problems with the Rasch Model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D., & Mead, R. The Use of Measurement Models in the Definition and Application of Social Science Variables. Final report to the Army Research Institute under Grant DAHC19-76-G-0011, June 1977.

APPENDIX A

USING THE CORNFIELD-TUKEY ALGORITHM FOR THE DERIVATION OF EXPECTED MEAN SQUARES FOR A x B x C x D DESIGN, A, B, D RANDOM EFFECTS, C FIXED

Source	i	j	k	m	Expected mean square[a]
A_i	$D_i = 1$	b	c	d	bcdA+cdAB+bcAD+cABD
B_j	a	$D_j = 1$	c	d	acdB+cdAB+acBD+cABD
C^*_k	a	b	$D_k = 0$	d	abdC+bdAC+adBC+adCD+dABC +bACD+aBCD+ABCD
D_m	a	b	c	$D_m = 1$	abcD+bcAD+acBD+cABD
AB_{ij}	$D_i = 1$	$D_j = 1$	c	d	cdAB+cABD
AC^*_{ik}	$D_i = 1$	b	$D_k = 0$	d	bdAC+dABC+bACD+ABCD
AD_{im}	$D_i = 1$	b	c	$D_m = 1$	cAD+cABD
BC^*_{jk}	a	$D_j = 1$	$D_k = 0$	d	adBC+dABC+aBCD+ABCD
BD_{jm}	a	$D_j = 1$	c	$D_m = 1$	acBD+cABD
C^*D_{km}	a	b	$D_k = 0$	$D_m = 1$	abCD+bACD+aBCD+ABCD
ABC^*_{ijk}	$D_i = 1$	$D_j = 1$	$D_k = 0$	d	dABC+ABCD
ABD_{ijm}	$D_i = 1$	$D_j = 1$	c	$D_m = 1$	cABD
AC^*D_{ikm}	$D_i = 1$	b	$D_k = 0$	$D_m = 1$	bACD+ABCD
BC^*D_{jkm}	a	$D_j = 1$	$D_k = 0$	$D_m = 1$	aBCD+ABCD
ABC^*D_{ijkm}	$D_i = 1$	$D_j = 1$	$D_k = 0$	$D_m = 1$	ABCD

$i = a,\ j = b,\ k = c,\ m = d$

$D_i = 1,\ D_j = 1,\ D_k = 0,\ D_m = 1$

Note. Adapted from Kirk (1968) and Winer (1971).

[a]For simplicity, capital letters in the EMS column represent subscripts of components, e.g., bcdA = bcd σ^2_A.

APPENDIX B

DERIVATION OF EXPECTED MEAN SQUARES AND QUASI-F RATIOS FOR COMPLETELY CROSSED DESIGN, A, B, D RANDOM FACTORS, AND C FACTOR FIXED

Source	Sources in E(MS)	Quasi-F ratio
A	A+AB+AD+ABD	$\frac{MS_A + MS_{ABD}}{MS_{AB} + MS_{AD}}$
B	B+AB+BD+ABD	$\frac{MS_B + MS_{ABD}}{MS_{AB} + MS_{BD}}$
C	C+AC+BC+CD+ABC+ACD+BCD+ABCD	$\frac{MS_C + MS_{ABC} + MS_{ACD} + MS_{BCD}}{MS_{AC} + MS_{BC} + MS_{CD} + MS_{ABCD}}$
D	D+AD+BD+ABD	$\frac{MS_D + MS_{ABD}}{MS_{AD} + MS_{BD}}$
AB	AB+ABD	$\frac{MS_{AB}}{MS_{ABD}}$
AC	AC+ABC+ACD+ABCD	$\frac{MS_{AC} + MS_{ABCD}}{MS_{ABC} + MS_{ACD}}$
AD	AD+ABD	$\frac{MS_{AD}}{MS_{ABD}}$
BC	BC+ABC+ACD+ABCD	$\frac{MS_{BC} + MS_{ABCD}}{MS_{ABC} + MS_{BCD}}$
BD	BD+ABD	$\frac{MS_{BD}}{MS_{ABD}}$
CD	CD+ACD+BCD+ABCD	$\frac{MS_{CD} + MS_{ABCD}}{MS_{ACD} + MS_{BCD}}$

APPENDIX B (Continued)

Source	Sources in E(MS)	Quasi-F ratio
ABC	ABC+ABCD	$\frac{MS_{ABC}}{MS_{ABCD}}$
ABD	ABD	Untestable
ACD	ACD+ABCD	$\frac{MS_{ACD}}{MS_{ABCD}}$
BCD	BCD+ABCD	$\frac{MS_{BCD}}{MS_{ABCD}}$
ABCD	ABCD	Untestable

APPENDIX C

DERIVATION OF DEGREES OF FREEDOM FOR NUMERATOR AND DENOMINATOR FOR QUASI-F RATIOS

The following algorithm can be used to derive the values for the degrees of freedom in the numerator and denominator for each of the sources:

$$\frac{(\text{sum of mean square terms in numerator of quasi-F ratio})^2}{(\text{sum of each mean square term in num. squared divided by its d.f.})};$$

and a similar relationship holds for denominator degrees of freedom, except that the terms are taken from the denominator of the quasi-F ratio instead of from the numerator.

A computational example for the "persons" (A) effect will be used to illustrate this procedure.

$$\frac{(MS_A + MS_{ABD})^2}{\frac{MS_A^2}{df_A} + \frac{MS_{ABD}^2}{df_B}} = \frac{(12.8 + .97)^2}{\frac{12.8^2}{236} + \frac{.97^2}{472}} = 272 \text{ degrees of freedom for the}$$

numerator of the A effect's quasi-F ratio.

$$\frac{(MS_{AB} + MS_{AD})^2}{\frac{MS_{AB}^2}{df_{AB}} + \frac{MS_{AD}^2}{df_{AD}}} = \frac{(1.05 + 2.45)^2}{\frac{1.05^2}{236} + \frac{2.45^2}{472}} = 705 \text{ degrees of freedom for the}$$

denominator of the A effect's quasi-F ratio.

The same procedure applies to computing the degrees of freedom for all of the other sources of variance.

APPENDIX D

TUKEY POST-HOC TEST

For F'-type quasi-F ratios, the numerator = Mean Square of Effect.

The following steps are required to obtain the components in the denominator:

1. Write the components of E(MS) effect (MS_C, in this case) in the numerator.

2. The first denominator term = MS(first term in E(MS) effect which is not part of the tested effect; e.g., not C in this case).

$$\frac{(C+AC+BC+CD+ABC+ACD+BCD+ABCD)}{\begin{array}{l}(AC+ABC+ACD+ABCD)+(BC+ABC+BCD+ABCD)+(CD+ACD+BCD+ABCD)\\ -(ABC+ABCD)-(ACD+ABCD)-(BCD+ABCD)+ABCD\end{array}} .$$

3. Write the E(MS) terms for denominator term #1.

4. Cross out common numerator and denominator terms.

5. Repeat steps 2, 3, 4 for the remaining terms in the numerator.

6. Subtract leftover terms in the denominator.

7. Cross out terms in the denominator.

8. Repeat steps 6 and 7.

9. If any terms are left over, add the appropriate Mean Square to the denominator.

$$F' = \frac{MS_C}{MS_{AC}+MS_{BC}+MS_{CD}-MS_{ABC}-MS_{ACD}-MS_{BCD}+MS_{ABCD}}$$

$$d.f. = \frac{(MS_{AC}+MS_{BC}+MS_{CD}-MS_{ABC}-MS_{ACD}-MS_{BCD}+MS_{ABCD})^2}{\text{Mean Square of each of the above terms squared, divided by its d.f.}} .$$

$$d.f. = \frac{(1.9012+2.2644+4.3081-.9093-1.1366-.6802+.8222)^2}{\frac{1.9012^2}{1652}+\frac{2.2644^2}{7}+\frac{4.3081^2}{14}+\frac{.9093^2}{1652}+\frac{1.1366^2}{3304}+\frac{.6802^2}{14}+\frac{.8222^2}{3304}} = 20.607.$$

The appropriate error term therefore equals:

$$1.9012 + 2.2644 + 4.3081 - .9093 - 1.1366 - .6802 + .8222 = 6.5698.$$

The value of the Tukey Q ratio for comparing MPFQC tables 4 vs. 6:

$$\frac{8.37 - 6.79}{\sqrt{6.5698/1422}} = 23.245.$$

The value of the Tukey Q ratio for comparing the average number of hits scored on tables 1 through 4 vs. the average for tables 5 through 8 is:

$$\frac{8.95 - 6.44}{\sqrt{6.5698/1422}} = 39.93.$$

The tabled Q value for statistical significance at the $p \leq .01$ level is 5.84, or $Q_{.01,\ 8,\ 20} = 5.84$. Thus, each of the above comparisons are highly statistically significantly different; more hits were scored on table 6 than on table 4; and more hits were scored on the average of tables 5 through 8 than on the average for tables 1 through 4.

APPENDIX E

MAGNITUDES OF EFFECTS, OR PROPORTION OF VARIANCE ACCOUNTED FOR BY EACH SOURCE (EFFECT)

In order to compute the value of the variance components for each source of variance (experimental "effect"), first note that the estimated value of each expected mean square (E(MS)) is the observed Mean Square term in Table 1. The calculations simply require that the expression for each E(MS) (see Appendix A) be equated to the appropriate values of the mean square. The resulting set of equations are then solved.

The general form for each equation is:

$$\text{variance of effect} = \frac{\text{MS effect} - \text{sum of the variance components times their number of levels for all other terms in the E(MS) effect}}{\text{number of levels in that effect}}$$

These calculations can be simplified if the most complex interaction term is done first (ABCD, in this experiment), and you then work your way up to the main effect terms.

To calculate the proportion of variance for each source, first sum the values for each component to find *the total variance*. The proportion of variance for each source equals the value of its variance component divided by the total variance.

The following table shows how the variance components for each of the sources in the present experiment were computed.

Table E-1

Values for the Proportions of Variance Accounted for by Each Source

Source	Value of variance component, assuming that E(MS) = Mean Square	
A	(12.7987-.0854-1.4765-.9694)/48 =	.2139
B	(7.7018-.0854-0-.9694)/5688 =	.0012
C	(732.7079-.6775-1.3551-3.1715-.0871-.3144-0-.8222)/1422 =	.5107
D	(34.7514-1.4765-0-.9694)/3792 =	.0085
AB	(1.0548-.9694)/24 =	.0036
AC	(1.9012-.0871-.3144-.8222)/6 =	.1129
AD	(2.4459-.9694)/16 -	.0923
BC	(2.2644-.0871-0-.8222)/711 =	.0019
BD	(.4012-.9694)/1896 =	0
CD	(4.3081-.3144-0-.8222)/474 =	.0067
ABC	(.9093-.8222)/3 =	.0290
ABD	.9694/8 =	.1212
ACD	(1.1366-.8222)/2 =	.1572
BCD[a]	(.6802-.8222)/237 =	0
ABCD	ABCD = .8222	.8222
Total		2.0813

Note. Dividing each value in this table by the total gives the proportion of variance listed for the effects (sources of variance) listed in Table 2 of the text.

[a]For example, $MS_{BCD} = a\,\sigma^2_{BCD} + \sigma^2_{ABCD}$, so that

$$\sigma^2_{BCD} = \frac{MS_{BCD} - \sigma^2_{ABCD}}{a}.$$

APPENDIX F

TABLE DIFFICULTIES AND HIT PROBABILITIES COMPUTED FROM THE RASCH LOGISTIC MODEL (WRIGHT, 1977)

Table	Difficulty, D	Probability of hit, assuming mean ability of 1.60	Probability of hit, assuming 1 S.D. above mean, or 2.34	Probability of hit, assuming 1 S.D. below mean, or .86
1	.811	.688	.822	.512
2	.879	.677	.812	.495
3	1.261	.584	.746	.401
4	.756	.699	.830	.526
5	-.346	.875	.936	.770
6	-.206	.859	.927	.744
7	-.948	.927	.964	.859
8	-2.207	.978	.990	.956

APPENDIX G

THE INFORMATION VALUE IN A RESPONSE

Consider an examinee with ability "A" grappling with a test item of difficulty "D." If the examinee's ability is greater than the item's difficulty, then he should get the item correct. But if the item's difficulty is greater than the examinee's ability, then the item should "win" and the examinee will respond incorrectly to the item.

Supposing that the examinee has 10 units of ability and the item has five units of difficulty, then the examinee should get the item correct. If the examinee has eight units of ability and the item has six units of difficulty, he should still get the item correct, but less often, or with a lower probability of responding correctly, than when his ability more greatly exceeds the item's difficulty. This intuitive theory has been developed mathematically (Wright, 1977), and will now be briefly outlined.

There are two parameters in this model: person ability A, and item difficulty D. The ability of person "v" and the difficulty of item "i" are combined by forming their difference on the latent variable $A_v - D_i$. This difference describes the theoretical probability of what happens when examinee v pits his ability against item i: Either the person or the item will "win." Since this difference can range from negative to positive infinity but the probability must stay between zero and one, the difference is applied to the exponent of a base. Specifically, the difference is applied as the exponential to base e: $e^{(A_v - D_i)}$. The probability of a correct answer is then simply the ratio

$$e^{(A_v - D_i)}/1 + e^{(A_v - D_i)} .$$

It would be very laborious to perform the calculations by hand for groups of hundreds of examinees. A computer program developed by Wright and Mead (1977) was used to calculate the difficulty of each of the eight MPFQC tables, and the ability of examinees on the basis of their total number of hits. Appendix F shows the difficulty for each of the eight tables, and the ability of students' various total "raw" scores (number of hits) in increments of five. Appendix F shows the probability of getting a hit on each of the tables, for examinees whose ability level is average, and plus or minus one standard deviation from the average.

The probability values in this table can also be used as a very useful index of how much the examiner learns about an examinee when the examinee gets a hit or misses) on a particular table. Intuitively, it would seem that a very easy item would be about as easy for nonmasters to get correct as it would be for masters. For example, consider table 8,

where the distance to the target is only seven meters. Virtually every student gets that particular trial "correct" (scores a hit). So, how can we tell who can shoot well and who cannot, on the basis of scores only for table 8? Suppose that the student scored 188 hits out of 240 trials. This means that he has average ability. This average ability, B, is equal to 1.60, using the scale of this statistical model. And, using the same model, the difficulty, D, of table 1 equals .811. Then the probability of hitting the target in table 1 for a person with average ability may be expressed as

$$p(\text{hit on table 1 for ability} = 1.60) = \frac{e^{(1.60 - .811)}}{1 + e^{(1.60 - .811)}} = \frac{2.2}{3.2} = .688.$$

Now suppose that we observe a student whose total score is 158 hits, which therefore gives him an ability level of 1.60 - .74 = .86, using the scale for this model. (Note that the number of hits can be easily translated into the Ability Scale from the computer printout for these data.)

The item difficulty for table 1 remains the same, at .811. Then

$$p(\text{hit on table 1 for ability} = .86) = \frac{e^{(.86 - .811)}}{1 + e^{(.86 - .811)}} = \frac{1.05}{2.05} = .512.$$

Now consider a score which is one standard deviation above the mean, or 1.60 + .74 = 2.34. Thus, ability = 2.34, and the total number of hits is 210. Then

$$p(\text{hit on table 1 for ability} = 2.34) = \frac{e^{(2.34 - .811)}}{1 + e^{(2.34 - .811)}} = \frac{4.61}{5.61} = .822.$$

We can perform similar calculations for the remaining seven tables, assuming average scores, and total scores which are either one standard deviation above or below the average.

DISTRIBUTION

ARI Distribution List

4 OASD (M&RA)
2 HQDA (DAMI-CSZ)
1 HQDA (DAPE-PBR
1 HQDA (DAMA-AR)
1 HQDA (DAPE-HRE-PO)
1 HQDA (SGRD-ID)
1 HQDA (DAMI-DOT-C)
1 HQDA (DAPC-PMZ-A)
1 HQDA (DACH-PPZ-A)
1 HQDA (DAPE-HRE)
1 HQDA (DAPE-MPO-C)
1 HQDA (DAPE-DW)
1 HQDA (DAPE-HRL)
1 HQDA (DAPE-CPS)
1 HQDA (DAFD-MFA)
1 HQDA (DARD-ARS-P)
1 HQDA (DAPC-PAS-A)
1 HQDA (DUSA-OR)
1 HQDA (DAMO-RQR)
1 HQDA (DASG)
1 HQDA (DA10-PI)
1 Chief, Consult Div (DA-OTSG), Adelphi, MD
1 Mil Asst. Hum Res, ODDR&E, OAD (E&LS)
1 HQ USARAL, APO Seattle, ATTN: ARAGP-R
1 HQ First Army, ATTN: AFKA-OI-TI
2 HQ Fifth Army, Ft Sam Houston
1 Dir, Army Stf Studies Ofc, ATTN: OAVCSA (DSP)
1 Ofc Chief of Stf, Studies Ofc
1 DCSPER, ATTN: CPS/OCP
1 The Army Lib, Pentagon, ATTN: RSB Chief
1 The Army Lib, Pentagon, ATTN: ANRAL
1 Ofc, Asst Sect of the Army (R&D)
1 Tech Support Ofc, OJCS
1 USASA, Arlington, ATTN: IARD-T
1 USA Rsch Ofc, Durham, ATTN: Life Sciences Dir
2 USARIEM, Natick, ATTN: SGRD-UE-CA
1 USATTC, Ft Clayton, ATTN: STETC-MO-A
1 USAIMA, Ft Bragg, ATTN: ATSU-CTD-OM
1 USAIMA, Ft Bragg, ATTN: Marquat Lib
1 US WAC Ctr & Sch, Ft McClellan, ATTN: Lib
1 US WAC Ctr & Sch, Ft McClellan, ATTN: Tng Dir
1 USA Quartermaster Sch, Ft Lee, ATTN: ATSM-TE
1 Intelligence Material Dev Ofc, EWL, Ft Holabird
1 USA SE Signal Sch, Ft Gordon, ATTN: ATSO-EA
1 USA Chaplain Ctr & Sch, Ft Hamilton, ATTN: ATSC-TE-RD
1 USATSCH, Ft Eustis, ATTN: Educ Advisor
1 USA War College, Carlisle Barracks, ATTN: Lib
2 WRAIR, Neuropsychiatry Div
1 DLI, SDA, Monterey
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-WGC
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-MR
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-JF
1 USA Artic Test Ctr, APO Seattle, ATTN: STEAC-MO-ASL
1 USA Artic Test Ctr, APO Seattle, ATTN: AMSTE-PL-TS
1 USA Armament Cmd, Redstone Arsenal, ATTN: ATSK-TEM
1 USA Armament Cmd, Rock Island, ATTN: AMSAR-TDC
1 FAA-NAFEC, Atlantic City, ATTN: Library
1 FAA-NAFEC, Atlantic City, ATTN: Hum Engr Br
1 FAA Aeronautical Ctr, Oklahoma City, ATTN: AAC-44D
2 USA Fld Arty Sch, Ft Sill, ATTN: Library
1 USA Armor Sch, Ft Knox, ATTN: Library
1 USA Armor Sch, Ft Knox, ATTN: ATSB-DI-E
1 USA Armor Sch, Ft Knox, ATTN: ATSB-DT-TP
1 USA Armor Sch, Ft Knox, ATTN: ATSB-CD-AD
2 HQUSACDEC, Ft Ord, ATTN: Library
1 HQUSACDEC, Ft Ord, ATTN: ATEC-EX-E-Hum Factors
2 USAEEC, Ft Benjamin Harrison, ATTN: Library
1 USAPACDC, Ft Benjamin Harrison, ATTN: ATCP-HR
1 USA Comm-Elect Sch, Ft Monmouth, ATTN: ATSN-EA
1 USAEC, Ft Monmouth, ATTN: AMSEL-CT-HDP
1 USAEC, Ft Monmouth, ATTN: AMSEL-PA-P
1 USAEC, Ft Monmouth, ATTN: AMSEL-SI-CB
1 USAEC, Ft Monmouth, ATTN: C, Facl Dev Br
1 USA Materials Sys Anal Agcy, Aberdeen, ATTN: AMXSY-P
1 Edgewood Arsenal, Aberdeen, ATTN: SAREA-BL-H
1 USA Ord Ctr & Sch, Aberdeen, ATTN: ATSL-TEM-C
2 USA Hum Engr Lab, Aberdeen, ATTN: Library/Dir
1 USA Combat Arms Tng Bd, Ft Benning, ATTN: Ad Supervisor
1 USA Infantry Hum Rsch Unit, Ft Benning, ATTN: Chief
1 USA Infantry Bd, Ft Benning, ATTN: STEBC-TE-T
1 USASMA, Ft Bliss, ATTN: ATSS-LRC
1 USA Air Def Sch, Ft Bliss, ATTN: ATSA-CTD-ME
1 USA Air Def Sch, Ft Bliss, ATTN: Tech Lib
1 USA Air Def Bd, Ft Bliss, ATTN: FILES
1 USA Air Def Bd, Ft Bliss, ATTN: STEBD-PO
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: Lib
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: ATSW-SE-L
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: Ed Advisor
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: DepCdr
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: CCS
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCASA
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCACO-E
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCACC-CI
1 USAECOM, Night Vision Lab, Ft Belvoir, ATTN: AMSEL-NV-SD
3 USA Computer Sys Cmd, Ft Belvoir, ATTN: Tech Library
1 USAMERDC, Ft Belvoir, ATTN: STSFB-DQ
1 USA Eng Sch, Ft Belvoir, ATTN: Library
1 USA Topographic Lab, Ft Belvoir, ATTN: ETL-TD-S
1 USA Topographic Lab, Ft Belvoir, ATTN: STINFO Center
1 USA Topographic Lab, Ft Belvoir, ATTN: ETL-GSL
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: CTD-MS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATS-CTD-MS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TE
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TEX-GS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTS-OR
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTD-DT
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTD-CS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: DAS/SRD
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TEM
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: Library
1 CDR, HQ Ft Huachuca, ATTN: Tech Ref Div
2 CDR, USA Electronic Prvg Grd, ATTN: STEEP-MT-S
1 CDR, Project MASSTER, ATTN: Tech Info Center
1 Hq MASSTER, USATRADOC, LNO
1 Research Institute, HQ MASSTER, Ft Hood
1 USA Recruiting Cmd, Ft Sherdian, ATTN: USARCPM-P
1 Senior Army Adv., USAFAGOD/TAC, Elgin AF Aux Fld No. 9
1 HQ USARPAC, DCSPER, APO SF 96558, ATTN: GPPE-SE
1 Stimson Lib, Academy of Health Sciences, Ft Sam Houston
1 Marine Corps Inst., ATTN: Dean-MCI
1 HQUSMC, Commandant, ATTN: Code MTMT 51
1 HQUSMC, Commandant, ATTN: Code MPI-20
2 USCG Academy, New London, ATTN: Admission
2 USCG Academy, New London, ATTN: Library
1 USCG Training Ctr, NY, ATTN: CO
1 USCG Training Ctr, NY, ATTN: Educ Svc Ofc
1 USCG, Psychol Res Br, DC, ATTN: GP 1/62
1 HQ Mid-Range Br, MC Det, Quantico, ATTN: P&S Div

1 US Marine Corps Liaision Ofc, AMC, Alexandria, ATTN: AMCGS–F
1 USATRADOC, Ft Monroe, ATTN: ATRO–ED
6 USATRADOC, Ft Monroe, ATTN: ATPR–AD
1 USATRADOC, Ft Monroe, ATTN: ATTS–EA
1 USA Forces Cmd, Ft McPherson, ATTN: Library
2 USA Aviation Test Bd, Ft Rucker, ATTN: STEBG–PO
1 USA Agcy for Aviation Safety, Ft Rucker, ATTN: Library
1 USA Agcy for Aviation Safety, Ft Rucker, ATTN: Educ Advisor
1 USA Aviation Sch, Ft Rucker, ATTN: PO Drawer O
1 HQUSA Aviation Sys Cmd, St Louis, ATTN: AMSAV–ZDR
2 USA Aviation Sys Test Act., Edwards AFB, ATTN: SAVTE–T
1 USA Air Def Sch, Ft Bliss, ATTN: ATSA TEM
1 USA Air Mobility Rsch & Dev Lab, Moffett Fld, ATTN: SAVDL–AS
1 USA Aviation Sch, Res Tng Mgt, Ft Rucker, ATTN: ATST–T–RTM
1 USA Aviation Sch, CO, Ft Rucker, ATTN: ATST–D–A
1 HQ, DARCOM, Alexandria, ATTN: AMXCD–TL
1 HQ, DARCOM, Alexandria, ATTN: CDR
1 US Military Academy, West Point, ATTN: Serials Unit
1 US Military Academy, West Point, ATTN: Ofc of Milt Ldrshp
1 US Military Academy, West Point, ATTN: MAOR
1 USA Standardization Gp, UK, FPO NY, ATTN: MASE–GC
1 Ofc of Naval Rsch, Arlington, ATTN: Code 452
3 Ofc of Naval Rsch, Arlington, ATTN: Code 458
1 Ofc of Naval Rsch, Arlington, ATTN: Code 450
1 Ofc of Naval Rsch, Arlington, ATTN: Code 441
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Acous Sch Div
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Code L51
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Code L5
1 Chief of NavPers, ATTN: Pers-OR
1 NAVAIRSTA, Norfolk, ATTN: Safety Ctr
1 Nav Oceanographic, DC, ATTN: Code 6251, Charts & Tech
1 Center of Naval Anal, ATTN: Doc Ctr
1 NavAirSysCom, ATTN: AIR–5313C
1 Nav BuMed, ATTN: 713
1 NavHelicopterSubSqua 2, FPO SF 96601
1 AFHRL (FT) William AFB
1 AFHRL (TT) Lowry AFB
1 AFHRL (AS) WPAFB, OH
2 AFHRL (DOJZ) Brooks AFB
1 AFHRL (DOJN) Lackland AFB
1 HQUSAF (INYSD)
1 HQUSAF (DPXXA)
1 AFVTG (RD) Randolph AFB
3 AMRL (HE) WPAFB, OH
2 AF Inst of Tech, WPAFB, OH, ATTN: ENE/SL
1 ATC (XPTD) Randolph AFB
1 USAF AeroMed Lib, Brooks AFB (SUL–4), ATTN: DOC SEC
1 AFOSR (NL), Arlington
1 AF Log Cmd, McClellan AFB, ATTN: ALC/DPCRB
1 Air Force Academy, CO, ATTN: Dept of Bel Scn
5 NavPers & Dev Ctr, San Diego
2 Navy Med Neuropsychiatric Rsch Unit, San Diego
1 Nav Electronic Lab, San Diego, ATTN: Res Lab
1 Nav TrngCen, San Diego, ATTN: Code 9000–Lib
1 NavPostGraSch, Monterey, ATTN: Code 55Aa
1 NavPostGraSch, Monterey, ATTN: Code 2124
1 NavTrngEquipCtr, Orlando, ATTN: Tech Lib
1 US Dept of Labor, DC, ATTN: Manpower Admin
1 US Dept of Justice, DC, ATTN: Drug Enforce Admin
1 Nat Bur of Standards, DC, ATTN: Computer Info Section
1 Nat Clearing House for MH–Info, Rockville
1 Denver Federal Ctr, Lakewood, ATTN: BLM
12 Defense Documentation Center
4 Dir Psych, Army Hq, Russell Ofcs, Canberra
1 Scientific Advsr, Mil Bd, Army Hq, Russell Ofcs, Canberra
1 Mil and Air Attache, Austrian Embassy
1 Centre de Recherche Des Facteurs, Humaine de la Defense Nationale, Brussels
2 Canadian Joint Staff Washington
1 C/Air Staff, Royal Canadian AF, ATTN: Pers Std Anal Br
3 Chief, Canadian Def Rsch Staff, ATTN: C/CRDS(W)
4 British Def Staff, British Embassy, Washington
1 Def & Civil Inst of Enviro Medicine, Canada
1 AIR CRESS, Kensington, ATTN: Info Sys Br
1 Militaerpsykologisk Tjeneste, Copehagen
1 Military Attache, French Embassy, ATTN: Doc Sec
1 Medecin Chef, C.E.R.P.A.–Arsenal, Toulon/Naval France
1 Prin Scientific Off, Appl Hum Engr Rsch Div, Ministry of Defense, New Delhi
1 Pers Rsch Ofc Library, AKA, Israel Defense Forces
1 Ministeris van Defensie, DOOP/KL Afd Sociaal Psychologische Zaken, The Hague, Netherlands